RC25079 (C1011-011) November 22, 2010
Computer Science

IBM Research Report

Defining the Controlling Parameter in Constrained Discriminative Linear Transform for Supervised Speaker Adaptation

Danning Jiang[1], Dimitri Kanevsky[2], Emmanuel Yashchin[2], Yong Qin[1]

[1]IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

[2]IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

DEFINING THE CONTROLLING PARAMETER IN CONSTRAINED DISCRIMINATIVE LINEAR TRANSFORM FOR SUPERVISED SPEAKER ADAPTATION

Danning Jiang[1], *Dimitri Kanevsky*[2], *Emmanuel Yashchin*[2], *Yong Qin*[1]

[1]IBM China Research Lab, Beijing, China
[2]IBM Watson Research Center, NY, USA
{jiangdn, qinyong}@cn.ibm.com, {kanevsky, yashchi}@us.ibm.com

ABSTRACT

Constrained discriminative linear transform (CDLT) optimized with Extended Baum-Welch (EBW) has been presented in the literature as a discriminative speaker adaptation method that outperforms the conventional maximum likelihood algorithm. Defining the controlling parameter of EBW to achieve the best performance of speaker adaptation, however, still remains an open question. This paper presents an empirical study on this issue. Results of our experiment suggest that a log-linear relationship exists between the optimal controlling parameter and the amount of data. This relationship can be used to efficiently define the controlling parameter for each test speaker to improve CDLT performance. We also discuss the possibility of generalizing the log-linear rule to a wider range of learning problems because such knowledge can substantially reduce the computation effort for parameter tuning.

Index Terms**—** Extended Baum-Welch (EBW), Constrained Discriminative Linear Transform (CDLT), speaker adaptation, parameter tuning

1. INTRODUCTION

Speaker adaptation is critical for improving the performance of speech recognition systems when limited speaker data are available. Transform based adaptation methods like maximum likelihood linear regression (MLLR) [1] have proven efficient and effective for a variety of recognition tasks. Inside the MLLR family, constrained MLLR (CMLLR) can be applied at the feature end and thus is preferable in server-based speech recognition systems, where memory and disk space generally cannot afford duplications of the speaker-specific models necessary when model-space adaptations are employed.

In contrast to conventional CMLLR, constrained discriminative linear transform (CDLT) optimizes the speaker-specific transform using a discriminative criterion like maximal mutual information (MMI) or minimum phone error rate (MPE) and Extended Baum-Welch (EBW). Previous studies [2,3] have shown that discriminative linear transform can outperform the maximum likelihood baseline. Similar to the parameter update in discriminative acoustic model training (DT), a controlling parameter E is involved in the update of transform using EBW. In principle, E represents the learning speed. The smaller the E is, the faster the learning speed, and vice versa. In existing discriminative adaptation research [2,3], E was set to the same empirical value as that used in DT.

However, the setting of the controlling parameter in CDLT can be very different from that in DT for at least two reasons. Firstly, as speaker adaptation occupies application system resources, it requires a more aggressive learning speed for efficiency considerations. In DT, model parameters can be optimized gradually, in multiple iterations, so the learning speed can be set to a slower value. Secondly, the setting of E depends on the amount of data. A larger amount of data requires a smaller E to reduce the weight of the original parameters, while a smaller amount of data needs a larger E to improve the stability. In a speech system, the amount of adaptation data available for each speaker often varies, and using a fixed E as in DT for all speakers will lead to suboptimal accuracy.

In this paper, we investigate how the controlling parameter affects the performance of CDLT in supervised speaker adaptation experiments, and then propose a log-linear formula to define the parameter for each test speaker based on the experimental observations. We also discuss the possibility of generalizing our method to defining controlling parameters for other optimization models, providing a theoretical proof for ridge regression.

2. CONSTRAINED DISCRIMINATIVE LINEAR TRANSFORM

2.1. CDLT formula

Similar to standard CMLLR, CDLT also transforms both model means and variances with the same speaker-specific matrix, so it can be applied at the feature end:

$$\hat{o}(t) = Ao(t) + b = W\zeta(t) \tag{1}$$

where $W = [b^T \quad A^T]^T$ and $\zeta(t) = [1 \quad o(t)^T]^T$.

Given a discriminative objective function $F(\lambda)$, a weak-sense auxiliary function is defined and EBW is used to optimize the transform. According to [2], the sufficient statistics required to estimate the i-th row of the transform are as follows:

$$\beta = \sum_{j,m} \sum_{t} (\gamma_{jm}^{num}(t) - \gamma_{jm}^{den}(t)) + D_{jm} \tag{2}$$

$$\begin{aligned} G^{(i)} = \sum_{j,m} \frac{1}{\sigma_{jm}^{(i)2}} \Big(& \sum_{t} \gamma_{jm}^{num}(t)\zeta(t)\zeta(t)^T \\ & - \sum_{t} \gamma_{jm}^{den}(t)\zeta(t)\zeta(t)^T + D_{jm} Z_{jm} \Big) \end{aligned} \tag{3}$$

$$k^{(i)} = \sum_{j,m} \frac{\mu_{jm}^{(i)}}{{\sigma_{jm}^{(i)}}^{2}} (\sum_{t} \gamma_{jm}^{num}(t)\zeta(t) - \sum_{t} \gamma_{jm}^{den}(t)\zeta(t) + D_{jm} \begin{bmatrix} 1 \\ \tilde{\mu}_{jm} \end{bmatrix}) \quad (4)$$

where $\gamma_{jm}(t)$ is the posterior probability at time t for mixture component m of state j, and

$$Z_{jm} = \begin{bmatrix} 1 & \tilde{\mu}_{jm}^{T} \\ \tilde{\mu}_{jm} & \tilde{\Sigma}_{jm} + \tilde{\mu}_{jm}\tilde{\mu}_{jm}^{T} \end{bmatrix} \quad (5)$$

$$\tilde{\mu}_{jm} = A^{-1}(\mu_{jm} - b) \quad (6)$$

$$\tilde{\Sigma}_{jm} = A^{-1}\Sigma_{jm}{A^{-1}}^{T} \quad (7)$$

Then the linear transform can be estimated in the same way with CMLLR. If $\hat{w}_i$ denotes the i-th row of $\hat{W}$ and p_i is the corresponding extended cofactor row vector, then the linear transform can be estimated row-by-row by solving the following equation:

$$\beta \frac{p_i}{p_i w_i^{T}} - w_i G^{(i)} + k^{(i)} = 0 \quad (8)$$

2.2. The controlling parameter

As the discriminative transform is updated via EBW technique, a Gaussian-specific smoothing factor D_{jm} is required in the optimization. Theoretically, the smoothing factor should be a large constant to guarantee convergence of the parameter update procedure. It also controls learning speed of the optimization. The smaller the factor is, the faster the learning, and vice versa. D_{jm} is generally defined by setting it to $E\sum_{t}\gamma_{jm}^{den}(t)$, where E is the controlling parameter empirically set to a value inside [1.0, 2.0].

As discussed earlier, the controlling parameter E used in CDLT estimation can be very different from that used in DT. On the one hand, it should be aggressive enough to achieve a fast learning speed, so that the adaptation can be done efficiently. On the other hand, it still needs to be large enough to guarantee the convergence of optimization. The optimal setting of E also depends on the amount of adaptation data. A larger amount of data requires a smaller E to reduce the weight of the original parameters, while a smaller amount of data requires a larger E to improve stability of the parameter updates. In light of the above argument, the problem of defining the controlling parameter becomes an important research question that warrants further study.

Besides considerations of the learning speed, the setting of E also needs to guarantee that $G^{(i)}$ defined by (3) is invertible, for the inversion of $G^{(i)}$ is required in the transform estimation. Fortunately, this can be guaranteed in practical situations (details can be found in Appendix A).

3. EXPERIMENTS

We carried out supervised adaptation experiments to explore how the controlling parameter affects the CDLT performance. For the sake of simplicity, we only considered a single global transform. Boosted MMI [4] objective function was used in the CDLT estimation.

The experiments were based on an English speech recognition system. The acoustic model consists of 5k tied-states and 200k Gaussian components, trained on 2000 hours of data. The recognition features were 24-d vectors computed via an LDA+STC projection from 48-d MFCC features (the static cepstra plus the 1st, 2nd and 3rd order derivatives). SAT training was first performed on the features, where the speaker-specific transforms were estimated via CMLLR, and then feature-space and model-space MPE training was performed based on the SAT model. The language model used in the experiments was a general purpose trigram model.

Two sets of test data were used in the experiments, both of which were real user data collected from an English dictation system at different periods of time. Test set 1 was composed of 26 speakers, and test set 2 of 21 speakers. For each test speaker, separate adaptation data and test data sets were available. The adaptation data of each speaker was 4 minutes long, recorded by the speaker in the enrollment stage reading the prompts. The test data of each speaker was 7~20 minutes long, and it may have included various real-life background noises. Test set 1 was used to study the effects of the controlling parameter and for tuning, while test set 2 was purely for evaluating CDLT performance.

The discriminative adaptation followed the lattice-based framework. The denominator lattices were decoded using the un-adapted acoustic model and a unigram LM. Before the CDLT adaptation, CMLLR was first performed to initialize the discriminative transform.

3.1. Effects of the controlling parameter

Theoretically, the controlling parameter E represents the learning speed in optimization: the smaller E is, the faster the learning. Figure 1 shows the effects of E on WER for test set 1 by comparing the CDLT performance with E=0.2, 0.5, 1.0, 2.0 when the amount of adaptation data varies from 30 seconds to 4 minutes. It can be seen that when the adaptation data is relatively sufficient (i.e. >2 minutes), a smaller E (or faster learning speed) can lead to a lower WER, while for smaller amounts of data (i.e. <2 minnutes), an aggressive learning speed will cause the optimization to diverge, resulting in a very high WER. In the following experiments, we use E=0.5 as the CDLT baseline, for it corresponds to the fastest learning rate that can guarantee the convergence for all amounts of data.

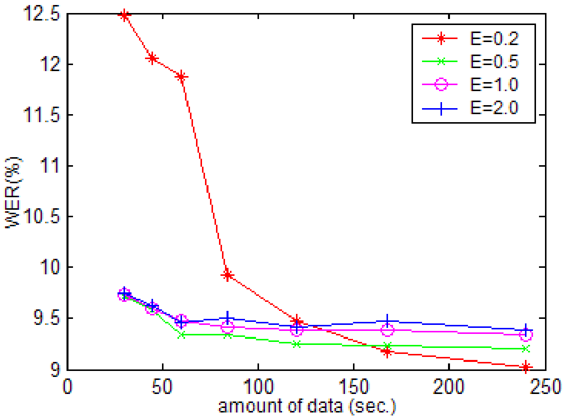

Figure 1. Comparison of CDLT performance with E=0.2, 0.5, 1.0, 2.0 when the amount of adaptation data varies (on test set 1)

Figure 2 shows the best E obtained via manual tuning for each amount of data on test set 1 (displayed as the solid line with asterisk marks). The plot shows that the optimal E of CDLT highly depends on the amount of adaptation data, and the dependence tends to be log-linear. The correlation coefficient of $\ln(E)$ and $\ln(n)$ (n denotes the amount of data) computed on test set 1 was -0.98, confirming the log-linear dependence. This relationship can be described by the following equation:

$$\ln E = 1.802 - 0.618 * \ln(n) \qquad (9)$$

where the linear parameters were estimated via linear regression based on test set 1. The dotted line in figure 2 shows the regressed E contour, which is rather close to the manually tuned one.

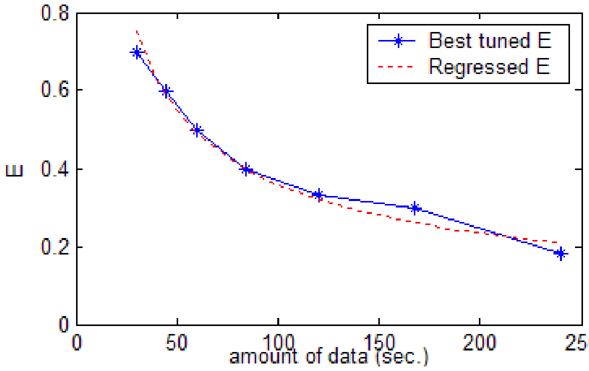

Figure 2. Comparison of the best tuned E and the regressed E when the amount of adaptation data varies (on test set 1).

3.2. Evaluation results

Tables 1 and 2 list the WERs of CDLT and CMLLR for test set 1 and test set 2, respectively. Three different controlling parameter setting methods were compared: a. E was set to 0.5 for all amounts of data (the baseline); b. E was manually tuned to obtain the best accuracy specifically for each amount of data; c. E was defined using (9).

Table 1. WERs of CDLT with different E setting methods and CMLLR on test set 1.

	CMLLR	CDLT		
		E=0.5	Tuned E	Predicted E
30 sec.	9.74	9.71	9.67	9.70
45 sec.	9.70	9.58	9.57	9.57
1.0 min.	9.52	9.34	9.34	9.34
1.4 min.	9.51	9.34	9.29	9.29
2.0 min.	9.50	9.25	9.16	9.16
2.8 min.	9.44	9.24	9.12	9.15
4.0 min.	9.45	9.21	9.02	9.03

For both test sets, CDLT with the E defined by (9) (denoted as predicted E) clearly outperformed the CDLT baseline (E=0.5) and CMLLR, and it performed similarly to CDLT when E was manually tuned. The WER was reduced more when there were relatively sufficient adaptation data (i.e. >=2 minutes). When 4 minutes adaptation data were used for each speaker on test sets 1 and 2, CDLT based on the predicted E reduced the WER by 4.4% and 3.0% compared to CMLLR, and reduced the WER by 2.0% and 2.2% compared to the CDLT baseline, respectively. Since the controlling parameter prediction formula was regressed only based on test set 1, the good performance on test set 2 suggests that the parameter defining method can be generalized to new test speakers.

Table 2. WERs of CDLT with different E setting methods and CMLLR on test set 2.

	CMLLR	CDLT		
		E=0.5	Tuned E	Predicted E
30 sec.	7.20	7.26	7.14	7.14
45 sec.	7.21	7.18	7.14	7.15
1.0 min.	7.13	7.12	7.08	7.12
1.4 min.	7.14	7.01	7.00	7.03
2.0 min.	7.11	7.00	6.95	6.96
2.8 min.	7.04	6.94	6.87	6.87
4.0 min.	7.03	6.97	6.82	6.82

We also submitted the data of test set 2 to paired t-test to check whether these improvements were statistically significant. The results revealed that when 4 minutes adaptation data were used, the improvements of CDLT based on the predicted E over both CMLLR and CDLT baseline were significant (*ps* <.05). When less adaptation data (>=2 min.) were used, the improvements over CMLLR were still significant (*ps* <.05), while the significance level of improvements over CDLT baseline gradually decreased (*p* =.076 for 2.8 minutes data and *p* =.138 for 2 minutes data).

4. DISCUSSIONS OF THE LOG-LINEAR RELATIONSHIP

In these experiments, we have found that the controlling parameter of EBW in CDLT estimation has a log-linear dependence on the amount of adaptation data. As controlling parameters are involved in many optimization models, an interesting question is whether such log-linear dependence exists for general classes of similar models. Knowledge of this kind of relationship would save substantial computational effort in the process of parameter tuning. We explore this question for the case of multiple regression. Since linear regression is the backbone of many problems in learning, the rule discovered on multiple regressions can be expected to hold for a wider range of situations.

Let the regression model be

$$Y = X\beta + \varepsilon \qquad (10)$$

where X is the design matrix (n by p), corresponding to n samples and p variables, $Y = \{y_1, y_2, ..., y_n\}$ are the response variables and $\varepsilon = \{e_1, e_2, ..., e_n\}$ are the noise terms. Ridge regression [5] is used instead of the conventional multiple regression to estimate the slopes $\beta = \{\beta_1, \beta_2, ..., \beta_P\}$ in order to achieve maximal predictive ability. It has been proven that a procedure for achieving the best predictive ability calls for estimation of β by a vector b obtained as a solution of the problem:

$$\text{Minimize} \quad (1/n) \| Y - Xb \|^2 + \lambda \| b \|^2 \qquad (11)$$

In this setting, λ is called the "tuning parameter", and it plays a role similar to the controlling parameter E in EBW. It can be proven that one can expect an asymptotic inverse relationship between λ and the sample size n (see appendix B), which implies the relationship of type (9) with a slope of -1. This means that the log-linear relationship can be widely used to predict the optimal tuning parameter for many learning problems that can be represented as instances of penalized linear regression.

It appears that slope -0.618 in (9) is not as aggressive as the asymptotic theory would suggest. While this could be caused by model differences, it may be also be related to practical issues like the specific features used in the speech recognition system, mismatch of noise conditions in the adaptation and test data, or presence of outliers. Further research is needed to explore inherent properties of EBW and how they are affected by practical issues.

5. CONCLUSIONS

In this paper, we presented an empirical study that investigates impacts of the EBW controlling parameter on the adaptation performance of constrained discriminative linear transform (CDLT). The experimental results suggest that a log-linear relationship exists between the optimal setting of the controlling parameter E and the amount of adaptation data, which could be used to define the controlling parameter for a test speaker. With E set based on the log-linear relationship, CDLT performance was better than the CDLT baseline (where E was set to a fixed value), and better than CMLLR. The improvements were more significant when sufficient adaptation data (>=2 minutes) were available.

We also discussed the value of investigating the log-linear relationship in other learning situations, since generalization of this finding can substantially reduce the computation effort related to parameter tuning. Specifically, we explored the case of ridge regression, and proved that the log-linear relationship does exist there as well. Based on these results, we can expect that the log-linear relationship holds more generally in multiple settings, since regularized linear regression is the backbone of many learning problems.

6. ACKNOWLEDGEMENTS

The authors sincerely thank Peder Olsen for the math discovering effects of the controlling parameter on the inversion of G matrix and Sara Basson for editing assistance.

7. APPENDIX

A. Inversion of G matrixes

To show how the controlling factor affects the inversion of $G^{(i)}$ defined by (3), we rewrite (3) as follows:

$$G^{(i)} = A + EB \tag{12}$$

where

$$A = \sum_{j,m} \frac{1}{\sigma_{jm}^{(i)2}} (\sum_t \gamma_{jm}^{num}(t)\zeta(t)\zeta(t)^T - \sum_t \gamma_{jm}^{den}(t)\zeta(t)\zeta(t)^T) \tag{13}$$

$$B = \sum_{j,m} \frac{1}{\sigma_{jm}^{(i)2}} \sum_t \gamma_{jm}^{den}(t) Z_{jm} \tag{14}$$

Then, $G^{(i)}$ is invertible since $\det(G^{(i)}) != 0$, and

$$\det(G^{(i)}) = \det(A + EB) = \det(B) * \det(B^{-1/2}AB^{-1/2} + EI) \tag{15}$$

As B is positive definite, $\det(B) > 0$. Denote $X = B^{-1/2}AB^{-1/2}$ and represent the eigenvalue decomposition as $X = P^T \Lambda P$, where Λ is the diagonal matrix of eigenvalues and P is the associated eigenvector matrix. Thus,

$$\begin{aligned}\det(B^{-1/2}AB^{-1/2} + EI) &= \det(P^T \Lambda P + EP^T P) \\ &= \det(P^T P) * \det(\Lambda + EI) = \prod_k (\lambda_k + E)\end{aligned} \tag{16}$$

From Eq. (16) and Eq. (15) we can find that as long as E is not equal to the negative of any of the eigenvalues of X, $G^{(i)}$ will be invertible. Fortunately, in our experiments E was almost never a zero point of (16), so the inversion of $G^{(i)}$ typically existed.

B. Ridge Regression

In what follows, we will show that λ in ridge regression can be represented as:

$$\ln(\lambda) = c0 - c1 * \ln(n) + o(\ln(n)) \tag{17}$$

In order to simplify the discussion, we introduce the parameter $d = \lambda * n$. We will show that for regression-type problems d tends to a constant as the sample size increases, implying a relationship of type (17).

In the case of multiple linear regression, one can show that the solution of (11) is of the form:

$$b = (X^T X + dI)^{-1} X^T$$

As shown in [5], a suitable d can be obtained by minimizing the Generalized Cross-Validation (GCV) criterion $V(d)$ defined by:

$$V(d) = (1/n) * \| (I - A(d))Y \|^2 / \{1 - TraceA(d)/n\}^2 \tag{18}$$

where

$$A(d) = X(X^T X + dI)^{-1} X^T . \tag{19}$$

For the limitation of space, we will only discuss the univariate case here. The generalizations for the multivariate case can be proven based on the Singular Value Decomposition (SVD) of X. In the univariate case, X consists of a single column, $X = \{x_1, x_2, \cdots, x_n\}^T$ and

$$A(d) = (\sum_{i=1}^n x_i^2 + d)^{-1} X^T X \tag{20}$$

$$b = (\sum_{i=1}^n x_i^2 + d)^{-1} (\sum_{i=1}^n x_i y_i) \tag{21}$$

The criterion $V(d)$ in this case can be represented as

$$V(d) = (1 - \frac{\sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 + d)})^{-1} [\frac{1}{n} \sum_{i=1}^n (y_i - bx_i)^2] \tag{22}$$

Minimization of $V(d)$ by d leads, after some algebra, to:

$$d^*(n) = (-\frac{1}{n} + r^2)^{-1} (\frac{1}{n} \sum_{i=1}^n x_i^2)(1 - r^2) \tag{23}$$

where r is the sample correlation coefficient. By (23), we obtain

$$\lim_{n -> \infty} d^*(n) = [p(1 - r^2)] / r^2 \tag{24}$$

thus proving (17).

8. REFERENCES

[1] M.J.F. Gales, “Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition,” *Computer Speech and Language*, 12(2), pp. 75-98, 1998.

[2] L. Wang and P.C. Woodland, “Discriminative Adaptive Training Using the MPE Criterion,” *ASRU 2003*, pp. 279-284.

[3] L. Wang and P.C. Woodland, “MPE-Based Discriminative Linear Transform for Speaker Adaptation,” *ICASSP 2004*, pp. 321-324.

[4] D. Povey, D. Kanevsky and B. Kingsbury, “Boosted MMI for Model and Feature-Space Discriminative Training,” *ICASSP 2008*, pp. 4057-4060.

[5] Golub,G., Heath, M. and Wahba, G. (1979). “Generalized Cross-validation as a method for choosing a good ridge parameter.” *Technometrics*, Vol. 21, pp. 215-223.